Testing Parametric versus Semiparametric Modelling in Generalized Linear Models[1]

Wolfgang HÄRDLE
Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin, Germany

Enno MAMMEN
Institut für Angewandte Mathematik
Ruprecht-Karls-Universität Heidelberg, Germany

Marlene MÜLLER
Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin, Germany

[1]The research for this paper was supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" at Humboldt University, Berlin (Germany). The work of M. Müller was supported in part by CentER, Tilburg University (The Netherlands). We thank Michael C. Burda for helpful discussion and comments on the economic applications. The paper is printed using funds made available by the Deutsche Forschungsgemeinschaft.

We consider a generalized partially linear model $E(Y|X,T) = G\{X^T\beta + m(T)\}$ where G is a known function, β is an unknown parameter vector, and m is an unknown function. The paper introduces a test statistic which allows to decide between a parametric and a semiparametric model: (i) m is linear, i.e. $m(t) = t^T\gamma$ for a parameter vector γ, (ii) m is a smooth (nonlinear) function. Under linearity (i) it is shown that the test statistic is asymptotically normal. Moreover, it is proved that the bootstrap works asymptotically. Simulations suggest that (in small samples) bootstrap outperforms the calculation of critical values from the normal approximation. The practical performance of the test is shown in applications to data on East–West German migration and credit scoring.

1 Introduction

In the analysis of discrete response variables one often models the expected value of the response as a nonlinear monotone function of a linear combination of the explanatory variables. Examples are Probit or Logit models where the nonlinear (link) function is the cumulative distribution function of a normal or logistic distribution, respectively, see McCullagh and Nelder (1989). Then the so-called *generalized linear model* has the form

$$E(Y|Z) = G(Z^T\theta) \tag{1.1}$$

with a known monotone function G and an unknown parameter θ. The model (1.1) combines computational feasibility (especially for discrete covariates) with good interpretability of the "index" $Z^T\theta$ and therefore has found wide application in all fields of applied statistics, see e.g. Fahrmeir and Tutz (1994), Maddala (1983). However, for some applications it may be argued that the assumption of linearity in (1.1) is too restrictive. Indeed it may be not even clear if the relationship between the influential variables and the response is monotone. A more complex relationship (allowing also for nonmonotone dependence) is given by the following semiparametric *generalized partially linear model*

$$E(Y|Z) = G\{X^T\beta + m(T)\} \tag{1.2}$$

where $Z = (X,T)$ is a split of Z into two components X and T, β is an unknown parameter and m is an unknown smooth function. For a discussion of model (1.2) and for further references, see Severini and Staniswalis (1994).

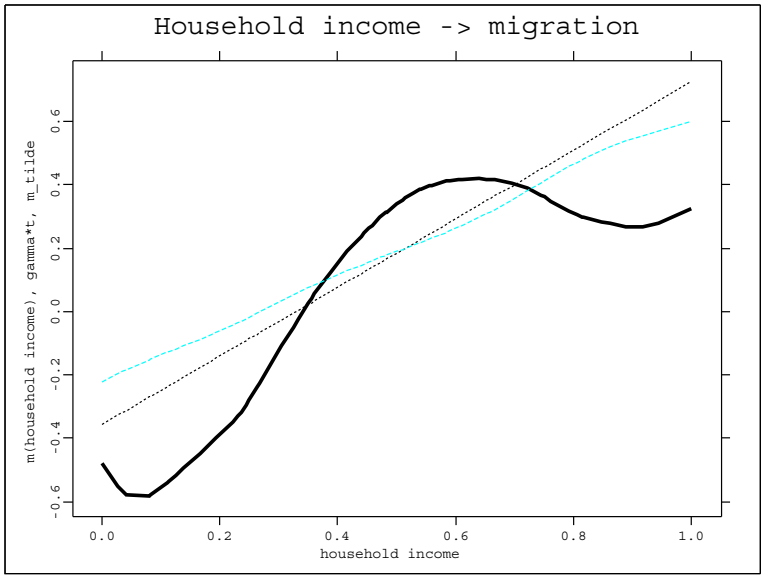

Figure 1: The influence $m(t)$ of household income (transformed to $[0,1]$) on migration intention. Nonparametric fit (thick black line), linear fit (thin black dashed line), and "biased" parametric estimate $\widetilde{m}$ (see (2.9), thin grey dashed line), $n = 402$.

As an example for a possible nonlinear dependence consider a model on East–West German migration in 1991 (data from the German Socio-Economic Panel for Mecklenburg-Vorpommern, a Land of the Federal State of Germany, GSOEP, 1991). The dependent variable is binary with $Y = 1$ (intention to move) or $Y = 0$ (intention to stay). As an explanatory variable serves besides some socioeconomic factors $X =$ (age, sex, friends in west, city size, unemployment) the variable $T =$ household income. Figure 1 shows a fit of the function m in the semiparametric model (1.2) using a logistic link function $G(u) = 1/\{1+\exp(-u)\}$. The estimated function is clearly nonlinear and shows a saturation in the intention to migrate for higher income housholds. The question is of course, whether the observed nonlinearity is significant.

In this paper we will discuss tests of the parametric hypothesis (1.1), i.e.

$$m(t) = t^T\gamma \quad \text{for a vector } \gamma, \tag{1.3}$$

versus the semiparametric alternative (1.2). Our tests indicate whether nonlinear shapes observed in nonparametric fits of m are significant. Furthermore, the pro-

posed tests complement the work of Severini and Staniswalis (1994), who consider estimation under model (1.2). Optimal rates for the nonparametric component and efficient estimation of the parametric component has been discussed in Mammen and van de Geer (1997). With identity link this model has been also analysed by Green (1987), Speckman (1983) and Robinson (1988). For a related model with semiparametric index see Carroll, Fan, Gijbels and Wand (1995). Most of the literature in this semiparametric context though was devoted to estimation and not to testing.

Our test is based on ideas of Hastie and Tibshirani (1990). [For a more general setup] they propose to apply the likelihood ratio test and to use χ^2 approximations for the calculation of critical values. Approximate degrees of freedom are derived by calculating expectation of asymptotic expansions of the test statistic under the null hypothesis. For this approach only heuristic justification has been given. We propose the following modifications of this approach.

First we correct for bias of nonparametric estimates. Secondly we modify the test statistic for the reason that different likelihoods [smoothed or unsmoothed likelihood, respectively] have been used in the calculation of the nonparametric or parametric component. For this modified test we can develop an asymptotic distribution theory. The test statistic has not an asymptotic χ^2 distribution. We propose to use bootstrap for the calculation of critical values and we can show that bootstrap works.

The next Section 2 introduces estimators of m, γ and β. These estimators will be used in the construction of the test statistics. The test and its asymptotic properties are discussed in Section 3. Section 4 reports on a small simulation study, the application to the migration example and another example on credit scoring. Remarks on the computation of the test statistics and proofs of our results are given in the appendix.

2 Estimation in the Parametric and in the Semiparametric Model

For the estimation of the parametric component β and the nonparametric component m we follow the approach of Severini and Staniswalis (1994). The method is based on quasi-likelihood estimation. The quasi-likelihood function is defined as

$$Q(\mu; y) = \int_{\mu}^{y} \frac{(s-y)}{V(s)}\, ds$$

where μ is the (conditional) expectation of Y, i.e. $\mu = G\{X^T\beta + m(T)\}$. It is assumed here that the conditional variance of Y is $\sigma^2 V(\mu)$ where σ is an unknown scale parameter and V is a known function. Quasi-likelihood functions are motivated by exponential families. Note that the maximum likelihood estimate $\widehat{\theta}$, based on an i.i.d. sample $Y_1, ..., Y_n$ from an exponential family, is given by

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} Q(\mu_i; Y_i) = 0.$$

In our model the quasi-likelihood function is given as

$$\mathcal{L}(m, \beta) = \sum_{i=1}^{n} Q(\mu_i; Y_i) \tag{2.1}$$

where $(Y_1, X_1, T_1), \ldots, (Y_n, X_n, T_n)$ is a sample of independent observations and $\mu_i = G\{X_i^T\beta + m(T_i)\}$. The parameter β is supposed to lie in $B \subset I\!R^p$. The covariates X_i, T_i are $I\!R^p$ and $I\!R^q$ valued. We assume that the response variable Y_i is real valued. Multidimensional responses can be treated similarly.

For the estimation of the nonparametric component m we make use of the following smoothed quasi-likelihood

$$\mathcal{L}^S(m(.), \beta) = \int \sum_{i=1}^{n} K_h(t - T_i)\; Q[G\{X_i^T\beta + m(t)\}; Y_i]\, dt, \tag{2.2}$$

where $K_h(u) = (h_1 \cdot \ldots \cdot h_q)^{-1} K(h_1^{-1}u_1, ..., h_q^{-1}u_q)$ is a kernel (defined on $I\!R^q$) with bandwidth (vector) $h = (h_1, ..., h_q)$. Following Severini and Staniswalis (1994),

Severini and Wong (1992) we put for $\beta \in B$

$$\widehat{m}_\beta = \arg\max_m \mathcal{L}^S(m, \beta), \tag{2.3}$$

$$\widehat{\beta} = \arg\max_\beta \mathcal{L}(\widehat{m}_\beta, \beta), \tag{2.4}$$

$$\widehat{m} = \widehat{m}_{\widehat{\beta}}. \tag{2.5}$$

In (2.3) minimization runs over functions $m(.)$. Because an integral is minimized by minimizing its integrand the value $\eta = \widehat{m}_\beta(t)$ is defined as the minimizer of the "local likelihood" $\sum_{i=1}^n K_h(t - T_i)\ Q[G\{X_i^T\beta + \eta\}; Y_i]$, see (2.2). Without loss of generality we always assume that the constant vector is not contained in the design space. An intercept is automatically modelled by the nonparametric component. Under this assumption the minimization in (2.3) and (2.4) is unique. For a discussion of these estimates see Severini and Staniswalis (1994).

Our test will be based on a comparison of the semiparametric estimates with the estimators $(\widetilde{\beta}, \widetilde{\gamma})$ in the parametric model

$$(\widetilde{\beta}, \widetilde{\gamma}) = \arg\max_{\beta,\gamma} \mathcal{L}^P(\gamma, \beta). \tag{2.6}$$

Here $\mathcal{L}^P(\gamma, \beta)$ is the quasi-likelihood function in model (1.1)

$$\mathcal{L}^P(\gamma, \beta) = \sum_{i=1}^n Q\{G(X_i^T\beta + T_i^T\gamma); Y_i\}. \tag{2.7}$$

The scale parameter σ can be estimated by

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\mu}_i)^2 / V(\widehat{\mu}_i), \tag{2.8}$$

where $\widehat{\mu}_i = G\{X_i^T\widehat{\beta} + \widehat{m}(T_i)\}$.

A direct comparison of $\widehat{m}(t)$ and $t^T\widetilde{\gamma}$ may be misleading because $\widehat{m}$ has a smoothing bias which is typically nonnegligible. This holds also if the hypothesis of linearity is true. To avoid this effect we will add to $t^T\widetilde{\gamma}$ a bias which will compensate for the bias of $\widehat{m}(t)$. This will be done by 'smoothing' of the function $t \to t^T\widetilde{\gamma}$. For this purpose we consider the artificial data set $\{\bar{Y}_i, X_i, T_i\} : i = 1, \ldots, n$ where $\bar{Y}_i = G(X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma})$ is the parametric fit of $E(Y_i|X_i, T_i)$. The function $\widetilde{m}$ is defined by the following smoothing step:

$$\widetilde{m} = \arg\max_m \int \sum_{i=1}^n K_h(t - T_i)\, Q[G\{X_i^T\widetilde{\beta} + m(t)\}; \bar{Y}_i]\, dt. \tag{2.9}$$

In the appendix we will show that under the hypothesis $\widetilde{m}(t)$ is asymptotically equivalent to $t^T\widetilde{\gamma}+$ the bias of $\widehat{m}(t)$. Therefore in the difference $\widehat{m}(t)-\widetilde{m}(t)$ the bias cancels asymptotically.

3 Testing the Parametric versus the Semiparametric Model

Our test procedures are based on the comparison of the parametric estimates $\widetilde{\beta},\widetilde{m}$ with the semiparametric estimates $\widehat{\beta},\widehat{m}$. A natural approach would be based on the likelihood ratio statistic $\mathcal{L}(\widehat{m},\widehat{\beta})-\mathcal{L}(\widetilde{m},\widetilde{\beta})$. Unfortunately, this test statistic does not work because in the construction of $\widehat{m}$ and $\widehat{\beta}$ two different likelihood functions (smoothed and unsmoothed) have been used. [A Taylor expansion of the test statistic, in particular of the i-th summand into $c_i\delta_i+d_i\delta_i^2$ with $\delta_i=X_i^T(\widehat{\beta}-\widetilde{\beta})+\widehat{m}(T_i)-\widetilde{m}(T_i)$, does not lead to a quadratic form.] This cannot be repaired by using the smoothed quasilikelihood $\mathcal{L}^S$ instead of $\mathcal{L}$.

We propose the following test statistic:

$$R_1=-2\sum_{i=1}^{n}Q(\widetilde{\mu}_i;\widehat{\mu}_i), \tag{3.1}$$

with $\widetilde{\mu}_i=G\{X_i^T\widetilde{\beta}+\widetilde{m}(T_i)\}$ and $\widehat{\mu}_i=G\{X_i^T\widehat{\beta}+\widehat{m}(T_i)\}$ for $i=1,\ldots,n$.

Note that for the case that the variance function V is constant R_1 is equal to $\sum_{i=1}^{n}(\widetilde{\mu}_i-\widehat{\mu}_i)^2/V$. In general, R_1 is equal to $\sum_{i=1}^{n}(\widetilde{\mu}_i-\widehat{\mu}_i)^2/V(\bar{\mu}_i)$, where $\bar{\mu}_i$ is a point between $\widetilde{\mu}_i$ and $\widehat{\mu}_i$. Therefore R_1 can be interpreted as a weighted quadratic deviation.

If the distribution of Y does not belong to an exponential family, the calculation of R_1 involves evaluation of n integrals. In these cases the following two modifications of R_1 are easier to compute. They are motivated by a Taylor expansion of R_1.

$$R_2 \;=\; \sum_{i=1}^{n}\frac{[G'\{X_i^T\widehat{\beta}+\widehat{m}(T_i)\}]^2}{V[G\{X_i^T\widehat{\beta}+\widehat{m}(T_i)\}]}\left\{X_i^T(\widehat{\beta}-\widetilde{\beta})+\widehat{m}(T_i)-\widetilde{m}(T_i)\right\}^2. \tag{3.2}$$

and

$$R_3 = \sum_{i=1}^{n} \frac{\{G'(X_i^T\widetilde{\beta}+T_i^T\widetilde{\gamma})\}^2}{V\{G(X_i^T\widetilde{\beta}+T_i^T\widetilde{\gamma})\}} \left\{X_i^T(\widehat{\beta}-\widetilde{\beta})+\widehat{m}(T_i)-\widetilde{m}(T_i)\right\}^2. \quad (3.3)$$

Theorem 3.1 discusses asymptotics of these test statistics. The test statistics are asymptotically equivalent on the null hypothesis and have an asymptotic normal distribution.

Theorem 3.1
Suppose that the assumptions (A1) - (A8) [see Section A2] apply. Then on the hypothesis $m_0(t) = t^T\gamma_0$, it holds that

(i) $R_1 = R_2 + o_p(v_n) = R_3 + o_p(v_n)$,

(ii) $v_n^{-1}(R_1 - e_n) \xrightarrow{D} N(0,1)$,

where e_n is a sequence with $e_n = h_{prod}^{-1}\int K(u)^2\,du\,\lambda_1 + O(h_{max}^2 h_{prod}^{-1})$ and v_n^2 is defined by $v_n^2 = 2h_{prod}^{-1}\int K^{(2)}(u)^2\,du\,\lambda_2$. Here we use the notation $h_{max} = max\{h_1,\ldots,h_q\}$ and $h_{prod} = h_1\cdot\ldots\cdot h_q$. The kernel $K^{(2)}$ is the convolution of K with itself. Furthermore,

$$\lambda_1 = E\frac{E\left[\frac{\sigma^2(X,T)G'(\eta)^2}{V^2(G(\eta))}|T\right]}{E\left[\frac{G'(\eta)^2}{V(G(\eta))}|T\right]}p(T)^{-1},$$

$$\lambda_2 = E\frac{E\left[\frac{\sigma^2(X,T)G'(\eta)^2}{V^2(G(\eta))}|T\right]^2}{E\left[\frac{G'(\eta)^2}{V(G(\eta))}|T\right]^2}p(T)^{-1}.$$

where $\sigma^2(X,T)$ is the conditional variance of Y, given (X,T) and where $\eta = X^T\beta_0 + T^T\gamma_0$. If the conditional variance $\sigma^2(X,T)$ is correctly specified by $\sigma^2 V(G(\eta))$ then λ_1 is equal to λ_2 and $\sigma^{-2}\lambda_1 = \sigma^{-2}\lambda_2$ is the Lebesgue measure of the support S_T of T.

Note in particular, that $\int K(u)^2 du \neq \int\{K^{(2)}(u)\}^2 du$. Therefore for the case that $\lambda_1 = \lambda_2$, Theorem 3.1 implies that a χ^2 approximation is not appropriate for

the distribution of R_1. The reason is that for kernel smoothing operators $\mathcal{K}$ it does not hold that $\mathcal{K}\mathcal{K} = \mathcal{K}$. This is in contrast to projection operators like B-splines, see Buja, Hastie and Tibshirani (1989). In particular, $\lambda_1 = \lambda_2$ holds if $Q(y;\mu)$ is the log-likelihood. Then R_1 is a modification of the log likelihood test.

For the asymptotic mean e_n an explicit formula can be given that contains conditional expectations of smoothed functions. Because it is rather lengthy it is omitted here.

Theorem 3.1 states that the test statistics R_1, R_2 and R_3 are asymptotically equivalent on the hypothesis. By standard arguments of asymptotic decision theory the asymptotic equivalence remains valid for contiguous alternatives (i.e. $n^{-1/2}$ neighbored alternatives). In a parametric setting this would imply that these three tests have asymptotic equivalent power. However, in our nonparametric setup the tests will have nontrivial power (power bounded away from the level and from 1) only for noncontiguous alternatives. Therefore, power functions may behave quite differently. A comparison of power functions based on simulations can be found in the next section.

3.1 Bootstrap tests

For two points s_n and t_n the nonparametric estimates $\widehat{m}(s_n)$ and $\widehat{m}(t_n)$ are asymptotically independent if the supports of the kernels $K_h(\bullet - s_n)$ and $K_h(\bullet - t_n)$ are disjoint. This may explain why, asymptotically, R_1 behaves approximately like a sum of $O(h_1^{-1} \cdot \ldots \cdot h_q^{-1})$ independent summands and has an asymptotic normal limitc. Because, typically, $h_1^{-1} \cdot \ldots \cdot h_q^{-1}$ is not very large, it can be suspected that normal approximations do not work well for R_1, see Härdle and Mammen (1993) for a related discussion. Therefore, for the calculation of quantiles, we advise not to use normal approximations. Instead, we propose to use the bootstrap. We discuss here three versions of bootstrap. The first version is Wild Bootstrap which is related to proposals of Wu (1986) [see also Beran (1986) and Mammen (1992)] and which was first proposed by Härdle and Mammen (1993) in nonparametric setups. Note that in our model the conditional distribution of Y is not specified besides (A1) and (A2).

The Wild Bootstrap procedure works as follows.

Step 1. Calculate residuals $\hat{\varepsilon}_i = Y_i - G(X_i^T\widehat{\beta} + \widehat{m}(T_i))$.

Step 2. Generate n i.i.d. random variables $\varepsilon_1^*, \ldots, \varepsilon_n^*$ with mean 0, variance 1 and which fulfill for a constant C that $|\varepsilon_i^*| \leq C$ (a.s.) for $i = 1, \ldots, n$.

Step 3. Put $Y_i^* = G(X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma}) + \hat{\varepsilon}_i\varepsilon_i^*$ for $i = 1, \ldots, n$.

Step 4. Calculate estimates $\widehat{\beta}^*, \widehat{m}^*, \widetilde{\beta}^*, \widetilde{\gamma}^*, \widetilde{m}^*$ based on the bootstrap samples (X_1, T_1, Y_1^*), ..., (X_n, T_n, Y_n^*). Furthermore, calculate test statistics R_1^*, R_2^* and R_3^*. The $(1-\alpha)$ quantiles of the distributions of R_1, R_2, and R_3 can be estimated by the $(1-\alpha)$ quantiles of the conditional distributions of R_1^*, R_2^* or R_3^*, respectively.

Under the additional model assumption

$$Var(Y|X,T) = \sigma^2 V(G(X^T\beta_0 + T^T\gamma_0))$$

we propose the following modification of the resampling. In Step 3 put $Y_i^* = G(X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma}) + \hat{\sigma}V\{G[X_i^T\widehat{\beta} + \widehat{m}(T_i)]\}^{1/2}\varepsilon_i^*$ for $i = 1, \ldots, n$ where $\hat{\sigma}^2$ is a consistent estimate of σ^2. In this case the condition that $|\varepsilon_i^*|$ is bounded can be weakened to the assumption that ε_i^* has subexponential tails, i.e. for a constant C it holds that $E(e^{[|\varepsilon_i^*|/C]}) \leq C$ for $i = 1, \ldots, n$ [compare (A2)].

In the special situation that $Q(y;\mu)$ is the log-likelihood (a semiparametric generalized linear model), the conditional distribution of Y_i is specified by $\mu_i = G(X_i^T\beta + T_i^T\gamma)$. Then we recommend to generate n independent $Y_1, \ldots, Y_n$ with distributions defined by $G(X_1^T\widetilde{\beta} + T_1^T\widetilde{\gamma}), \ldots, G(X_n^T\widetilde{\beta} + T_n^T\widetilde{\gamma})$, respectively. This is a version of parametric bootstrap. E.g. in the binary response example that we considered above, Y_i is a Bernoulli variable with parameter $\mu_i = G(X_i^T\beta + T_i^T\gamma)$. Hence, here it is reasonable to resample from the Bernoulli distribution with parameter $\widetilde{\mu}_i = G(X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma})$.

Theorem 3.2 shows that these three bootstrap procedures work (for their corresponding models).

Theorem 3.2
Suppose that the assumptions of Theorem 3.1 hold. In case of application of the second or third version of bootstrap assume that the just mentioned additional

model assumptions hold. Then it holds for $j = 1, 2, 3$, *that*

$$d_K(\mathcal{L}^*(R_j^*), \mathcal{L}(R_j)) \xrightarrow{P} 0$$

where $\mathcal{L}(R_j)$ *is the distribution of* R_j, $\mathcal{L}^*(R_j^*)$ *is the conditional distribution of* R_j^* *[given the sample], and* d_K *denotes the Kolmogorov distance, which is for two probability measures* μ *and* ν *(on the real line) defined as*

$$d_K(\mu, \nu) = \sup_{t \in I\!R} \Big| \mu(X \leq t) - \nu(X \leq t) \Big|.$$

Application of these three versions of bootstrap for β has been discussed in Mammen and van de Geer (1997). There the nonparametric component has been estimated by splines. The statement of the theorem does also hold if the residuals are defined as $\hat{\varepsilon}_i = Y_i - G(X_i^T \widetilde{\beta} + T_i^T \widetilde{\gamma})$. We have seen in our simulations for binary responses that the normal approximation in Theorem 3.1 (ii) is indeed inaccurate for small sample sizes, see Section 4, but that critical values are estimated quite well by bootstrap.

Our test statistic depends on the choice of the bandwidth h. Different values of h may lead to different observed significance levels, see Section 4. Small values of h have been motivated by asymptotic minimax theory , see Ingster (1993) and Lepski and Spokoiny (1995). In particular, the bandwidths proposed in these papers are of smaller order than optimal bandwidths for nonparametric estimation. However, it is difficult to adapt their abstract assumptions to practical settings.

We suggest to apply the test for different choices of h. Differences in observed critical values can be interpreted. Whereas test statistics with small choices of h look more for the appearance of wiggles of small length, large choices of h may detect better global deviances from linearity. So the inspection of the test statistic for different h gives an impression in which respect the function m differs significantly from linear functions.

3.2 Testing average linearity

In case that our test has rejected the hypothesis of linearity it may be of interest to get more insights about the reasons of the rejection. For the case of $d > 1$

we propose to test for average linearity in the direction of one covariate. For a given weight function $w(t_2, \ldots, t_q)$ with $\int w(t_2, \ldots, t_q) dt_2 \cdots dt_q = 1$ we consider the hypothesis that

$$\int m(t_1, \ldots, t_q) w(t_2, \ldots, t_q) dt_2 \cdots dt_q = \alpha + \beta t_1 \quad \text{for all } t_1 \text{ and for fixed } \alpha \text{ and } \beta. \tag{3.4}$$

Testing average linearity of m in t_1 is in particular appropriate in the following model. In this model it is assumed that there is no interaction term of t_1 and $(t_2, \ldots, t_q)$:

$$m(t_1, \ldots, t_q) = m_1(t_1) + m_{2,\ldots,q}(t_2, \ldots, t_q) \quad \text{for some functions } m_1,\ m_{2,\ldots,q}. \tag{3.5}$$

For a discussion of this additive model see Buja et al. (1989) and Hastie and Tibshirani (1990). In this model, hypothesis (3.4) reduces to

$$m_1(t_1) = \alpha + \beta t_1 \quad \text{for all } t_1 \text{ and for fixed } \alpha \text{ and } \beta. \tag{3.6}$$

Deviance from average linearity can be measured by the following test statistic

$$R_4 = \min_{a,b} \sum_{i=1}^{n} \frac{[G'\{X_i^T \widehat{\beta} + \widehat{m}(T_i)\}]^2}{V[G\{X_i^T \widehat{\beta} + \widehat{m}(T_i)\}]} \left\{\widehat{m}_1(T_i) - a - bT_i\right\}^2, \tag{3.7}$$

where $\widehat{m}_1(t_1) = \int \widehat{m}(t_1, \ldots, t_q) w(t_2, \ldots, t_q) dt_2 \cdots dt_q$. For the additive model (3.5), the nonparametric estimate $\widehat{m}_1$ of the additive component m_1 has been considered in Linton and Nielsen (1995), Tjøstheim and Auestad (1994), Chen, Härdle, Linton and Severance-Lossin (1996), and Fan, Härdle and Mammen (1995). In a modified definition, the "marginal integration" in the calculation of $\widehat{m}_1$ is replaced by a "marginal summation". For generalized additive models, asymptotics for the estimate $\widehat{m}_1$ is developed in Härdle, Huet, Mammen and Sperlich (1997). Furthermore a proof for asymptotic normality and consistency of bootstrap for the test statistic R_4 can be found there.

4 Simulations and Application

To verify the properties of our test procedure we have run a small simulation study. The following model was used to simulate data from a generalized (partially) linear

model

$$E(Y|X=x,T=t)=P(Y=1|x_1,x_2,t)=F\{2x_1+x_2+m(t)\},$$

where F is the standard logistic distribution function $F(u)=1/(1+e^{-u})$. X_1,X_2 and T are independent. X_1 and T have a uniform distribution on $[-1,1]$. The variable X_2 is discrete and takes five values in $[-1,1]$.

We performed simulations under the linearity hypothesis using $m(t)=t$. The sample size was $n=100$, 250 and 500 and the number of replications in the bootstrap resampling was $n^*=200$. The simulation results are based on 500 replications. For smoothing in this section the quartic kernel $K(u)=\frac{15}{16}(1-u^2)^2I(|u|\leq 1)$ was used.

Table 1 summarizes the results for $m(t)=t$. As can be seen bootstrap seems to work quite accurate for all three test statistics, at least for $\alpha\geq 0.05$.

As expected the normal approximation of Theorem 3.1 can be quite inaccurate for small sample sizes and it should not be used for the calculation of critical values of the test statistics R_1,R_2,R_3. This can be seen from Table 2.

The values in Table 2 concern only the tail of the distributions of R_1,R_2, and R_3 and of the normal limit, given in Theorem 3.1. In the central region there are much larger differences between the distributions of R_1,R_2, and R_3 and the normal limit, given in Theorem 3.1, as can be seen in Figure 2. There, density estimates for R_1,R_2,R_3 [using the 500 Monte Carlo replications under the linear model $m(t)=t$] are plotted together with the limiting normal density. The normal limit and the distributions of the test statistics are nearly separated.[The density estimates for R_1,R_2,R_3 are kernel estimates with bandwidth according to Silverman's rule of thumb, i.e. $h=1.06\cdot 2.62\cdot\widehat{\sigma}\cdot n^{-1/5}$ for the quartic kernel. For better comparison, the normal density has been analogously convoluted with a quartic kernel.] Similar plots can be found in Härdle and Mammen (1993) where a related test statistic has been discussed for testing parametric versus nonparametric regression.

Finally we have run our simulations with a function m consisting of a convex combination of the linear function $m(t)=t$ and the nonlinear function $m(t)=\cos(\pi t)$. Figure 3 shows the power functions of R_1 for these alternatives (black lines). The power has been plotted for four different significance levels. The power functions for R_2 and R_3 are almost the same and therefore they have been omitted.

α	0.01	0.05	0.10	0.15	0.20
LR(p)	0.010	0.070	0.138	0.190	0.248
LR(sp)	0.014	0.088	0.220	0.328	0.428
R_1	0.010	0.052	0.116	0.178	0.246
R_2	0.010	0.052	0.116	0.184	0.250
R_3	0.012	0.052	0.116	0.178	0.244
	$n = 100$, $h = 0.6$				
LR(p)	0.012	0.044	0.098	0.148	0.194
LR(sp)	0.020	0.080	0.158	0.234	0.322
R_1	0.020	0.052	0.094	0.138	0.180
R_2	0.020	0.052	0.096	0.136	0.184
R_3	0.022	0.052	0.094	0.142	0.182
	$n = 250$, $h = 0.5$				
LR(p)	0.008	0.046	0.094	0.140	0.198
LR(sp)	0.020	0.092	0.164	0.256	0.338
R_1	0.020	0.056	0.104	0.160	0.212
R_2	0.020	0.056	0.104	0.166	0.214
R_3	0.022	0.054	0.104	0.158	0.212
	$n = 500$, $h = 0.4$				

Table 1: Relative number of rejections for the test statistics R_1, R_2 and R_3 using the bootstrap method with $n^* = 200$. Compared with relative number of rejections for parametric LR statistic (=LR(p)) and semiparametric LR statistic using approximative degrees of freedom (=LR(sp)). 500 Monte Carlo replications.

The dashed lines in Figure 3 show (simulated) power functions for a parametric likelihood–ratio test LR_p. The hypothesis "$m(x,t) = F\{c + x\beta + t\gamma\}$ for some β and γ" is tested against the alternative: "$m(x,t) = F\{c + x\beta + t\gamma + \omega\cos(\pi t)\}$ for some c, β, γ and ω". In this setup R_1 achieves nearly the power of the parametric test LR_p. In other models we observed larger losses.

For comparison we have also included a likelihood ratio test LR_{sp} of the parametric against semiprametric hypothesis. Critical values have been calculated using χ^2 approximations and the definition of approximative degrees of freedom

α	0.01	0.05	0.10	0.15	0.20
R_1	0.000	0.002	0.010	0.012	0.020
R_2	0.000	0.000	0.006	0.010	0.012
R_3	0.002	0.001	0.014	0.024	0.030
	$n = 100$, $h = 0.6$				
R_1	0.000	0.008	0.018	0.026	0.032
R_2	0.000	0.006	0.016	0.022	0.028
R_3	0.000	0.014	0.020	0.030	0.038
	$n = 250$, $h = 0.5$				
R_1	0.004	0.010	0.016	0.028	0.034
R_2	0.004	0.010	0.016	0.026	0.032
R_3	0.006	0.012	0.020	0.030	0.036
	$n = 500$, $h = 0.4$				

Table 2: Relative number of rejections using normal approximations. 500 Monte Carlo replications.

of (Hastie and Tibshirani, 1990). A more detailed description of this test can be found in Müller (1997). The grey curves in Figure 3 show the power of this test. It achieves a similiar power as R_1. However it does not hold the nominal significance level under the hypothesis, see Table 1.

Let us now return to our introductory example on East–West German migration. Our interest in this subject has been inspired by an analysis of Burda (1993). His paper considers a sample of 3710 East Germans, which have been surveyed in 1991 in the German Socio-Economic Panel, see GSOEP (1991). Among other questions the East German participants have been asked, if they can imagine to move to the Western part of Germany or West Berlin. As in Burda's study we give the value 1 for those who responded positive and 0 if not. The economic model is based on the idea that a person will migrate if its utility (wage differential) will exceed the costs of migration. Of course, neither variable, wage differential or costs, is directly available. Hence proxy variables need to be used. The original data set of Burda (1993) contains 34 explanatory variables, with four of them continuous (age, income, rent, job tenure). The remaining variables are dummy variables (sex, partner, homeowner, family/friends in west, and further variables

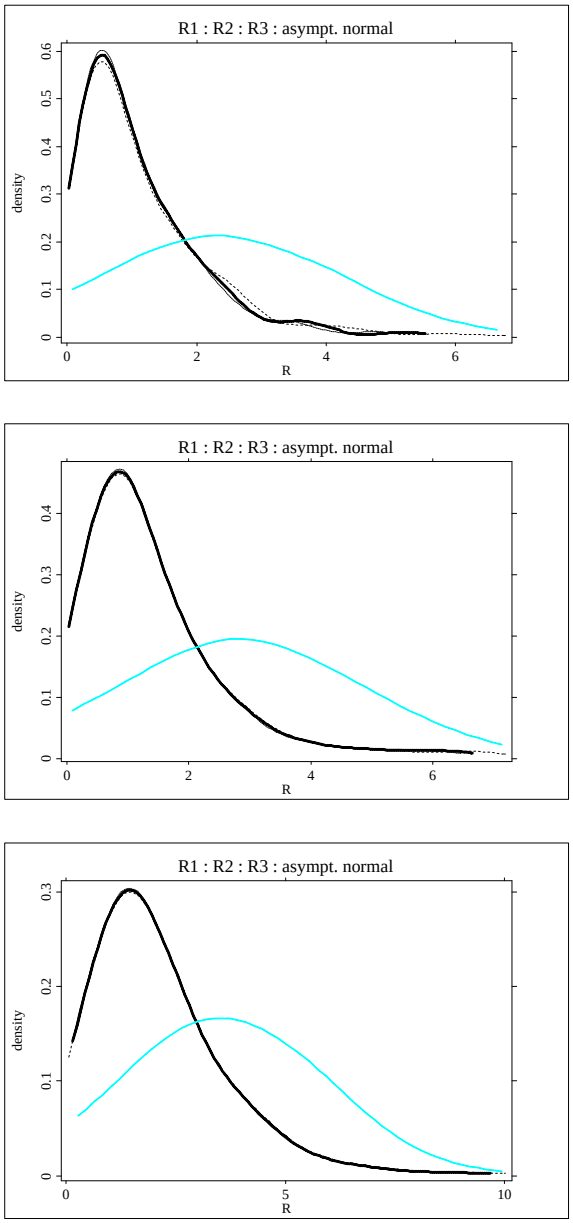

Figure 2: Density estimates for R_1 (thick line), R_2 (thin solid line), R_3 (thin dashed line) and normal density (grey line). $n = 100, 250, 500$ (upper to lower plot).

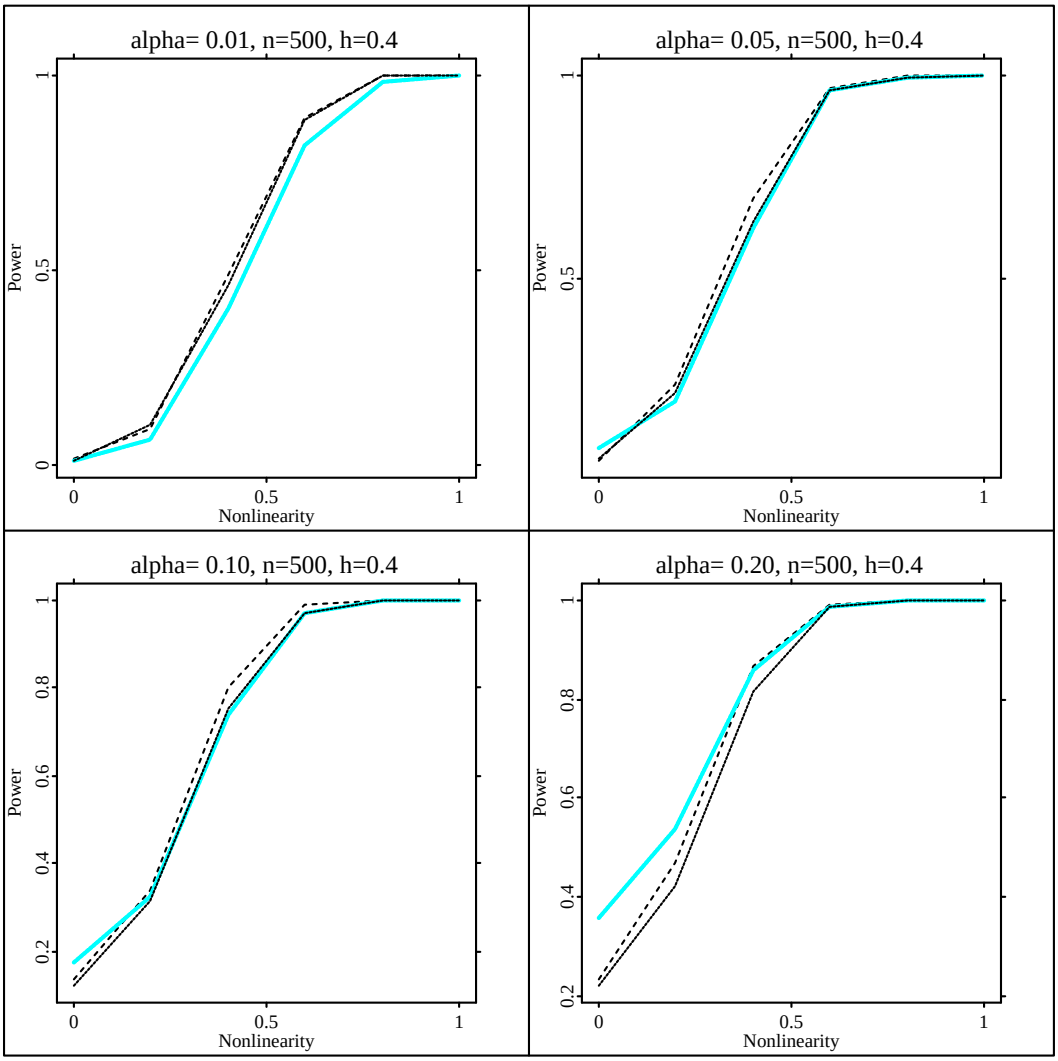

Figure 3: Power functions of test R_1 for $\alpha = 0.05, 0.10, 0.15, 0.20$ (black solid lines). $x, t \in [-1, 1]$ and $m(t) = (1-\nu)t + \nu\cos(\pi t)$, $\nu \in [0, 1]$, $n = 500$, $h = 0.4$. Compared to the power of the parametric LR test LR_p (dashed lines) and power of the semiparametric LR test LR_{sp} using approximate degrees of freedom (grey lines).

on occupation, city size, region, education).

		Yes	No	(in %)	
Y	**migration intention**	39.9	60.1		
X_1	**family/friends in west**	88.8	11.2		
X_2	**unemployed/job loss certain**	21.1	78.9		
X_3	**city size 10,000-100,000**	35.8	64.2		
X_4	**female**	50.2	49.8		
		Min	Max	Mean	S.D.
X_5	**age (years)**	18	65	39.93	12.89
T	**houshold income (DM)**	400	4000	2262.22	769.82

Table 3: Descriptive statistics for migration data. Sample from Mecklenburg–Vorpommern, $n = 402$.

It turns out, that regional variables have an important impact on the responses. For instance, the estimation is particularly difficult for East Germans living in East Berlin, since obviously other reasons may influence the intention to migrate than only the wage differential compared to costs. Also, the variables, which are most important, differ slightly between the five Eastern German states (plus East Berlin). Unemployment, for example, plays a stronger role in the Northern, less industrialized part of East Germany. In the following we give the estimation results for Mecklenburg–Vorpommern (in the very North of Eastern Germany) which leads to a sample size of $n = 402$. We have summarized some descriptive statistics in Table 3.

Table 4 shows the results of a logit fit, using a subset of covariates which have been chosen previously by a model selection procedure based on logit models. For simplicity both continuous variables (age, household income) have been linearly transformed to $[0, 1]$. The migration intention is definitely determined by age. However, also unemployment, city size and household income are highly significant.

A further analysis of this data set by a generalized additive model (keeping the logit link, but generalizing the influence of the age and income variables to nonparametric functions) showed that the age has a nearly perfect linear influence. Because of this relation, we modelled only the influence of household income as a nonparametric function. The coefficients for the parametric covariates are given

	Coeff.	(t–value)	Coeff.	(t–value)
const.	-0.358	(-0.68)	—	—
family/friends in west	0.589	(1.54)	0.599	(1.56)
unemployed/job loss certain	0.780	(2.81)	0.800	(2.87)
city size 10,000-100,000	0.822	(3.39)	0.842	(3.47)
female	-0.388	(-1.68)	-0.402	(-1.73)
age	-3.364	(-6.93)	-3.329	(-6.86)
household income	1.084	(1.90)	—	—
	Linear (logit)		Part. Linear	

Table 4: Logit coefficients and coefficients in a generalized partially linear model for migration data. Sample from Mecklenburg–Vorpommern, $n = 402$, $h = 0.3$.

in Table 4. The resulting fit $\widehat{m}$ (using bandwidth $h = 0.3$) for the function m is that shown in Figure 1 together with the linear fit (thin black dashed line) and the "biased" parametric fit $\widetilde{m}$ (see (2.9, thin grey dashed line). Recall that the estimate $\widetilde{m}$ is expected to be approximately equal to the sum of the parametric estimate and the bias of $\widehat{m}$.

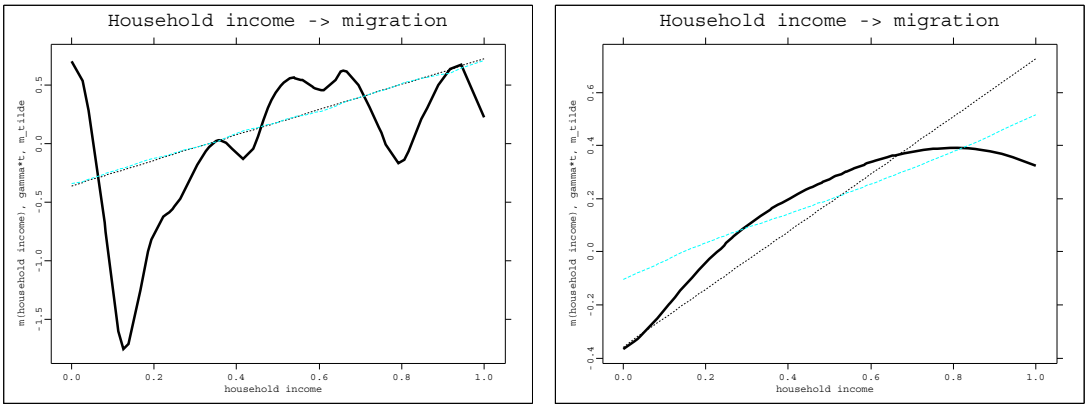

Figure 4: The influence $m(t)$ of household income on migration intention. Nonparametric fit (thick black line), linear fit (thin black dashed line), and "biased" parametric estimate $\widetilde{m}$ (thin grey dashed line). $n = 402$, bandwidths $h = 0.1$ (left) and $h = 0.5$ (right).

In Figure 4 we show the functions $\widehat{m}$ and $\widetilde{m}$ (together with the linear fit) for bandwidths $h = 0.1$ and $h = 0.5$. The nonparametric estimate $\widehat{m}$ in the migration example seems to be an obvious nonlinear function. However, it is difficult to judge the significance of the nonlinearity. In general, it cannot be excluded that the difference between the nonparametric and the linear fit may be caused by boundary and bias problems of $\widehat{m}$.

h	0.1	0.2	0.3	0.4	0.5
R_1	0.150	0.065	0.042	0.045	0.062
R_2	0.122	0.067	0.042	0.045	0.062
R_3	0.217	0.075	0.042	0.042	0.065
LR (sp)	0.053	0.066	0.048	0.035	

Table 5: Observed significance levels for linearity test for migration data, $n = 402$. $n^* = 400$ bootstrap replications.

Table 5 shows the results of the application of our tests from Section 3. The number of bootstrap simulations is always chosen as $n^* = 400$. We observe that all three tests R_1, R_2 and R_3 show nearly the same behaviour. The observed significance levels are given for different choices of the bandwidth h. Linearity is rejected (at 5% level) only for bandwidths $0.3, 0.4$. The different behaviour of the test for different h give some indication on possible deviance of m from linear functions. The appearance of wiggles of small length is not significant, see Figure 4 (left panel). However, the global shape of m seems to be not well approximable by linear functions. This result is in accordance with the estimate in Figure 1 and Figure 4 (right panel), where a saturation of the intention to migrate appears for the upper third of the data.

At the end of this section we will shortly present the application of our test statistic in a binary choice regression with a two-dimensional nonparametric function m. The data are a subsample from a training dataset on credit scoring, see Fahrmeir and Tutz (1994) and Fahrmeir and Hamerle (1984). The interest consists in finding how some factors are related to credit worthiness. We used the subsample of loans for cars, which has a sample size of $n = 284$ out of 1000. Some descriptive statistics for this subsample and a selection of covariates can be found in Table 6. The covariate "previous credit o.k." indicates that previous loans were paid without problems or that there were no previous loans. The variable

"employed" takes value 1 if the person taking the loan is employed with the same employer for at least one year. In the following statistical analysis we took logarithms of "amount" and "age" and transformed these values linearly to the interval $[0, 1]$.

		Yes	No	(in %)	
Y	**credit worthy**	73.6	26.4		
X_1	**previous credits o.k.**	66.2	33.8		
X_2	**employed**	73.2	26.8		
		Min	Max	Mean	S.D.
X_4	**duration (months)**	4	54	21.75	10.55
T_1	**amount (DM)**	428	14179	3902.31	2621.95
T_2	**age (years)**	19	75	34.16	10.81

Table 6: Descriptive statistics for credit data. Sample for credits for cars, $n = 284$.

A parametric logit model leads to the parameter estimates listed in Table 7. The influence of employment, duration and amount of credit have the expected sign. The negative influence of "previous credits o.k." is a bit astonishing, but may be explained that also people without previous loan fall in this category. The age variable shows a (global) positive influence in the logit fit, this will change together with the amount variable in the semiparametric fit. Note also, that both coefficients for "amount" and "age" are not significant at 10% level.

In a next step we fitted a generalized partially linear model to the data. Influence of "amount" and "age" has been fitted nonparametrically. The other variables have been modelled as linear covariates. For "duration" this has been done because, typically, it is divisible by 6 months. Figure 5 shows a scatterplot of the two variables "amount" and "age" on the left panel and the two–variate estimate $\widehat{m}$ (using a bandwidth $h = 0.4$ in both dimensions) on the right panel. It is difficult to check $\widehat{m}$ graphically for significant deviances from linearity. The big peak of $\widehat{m}$ is caused by only a few observations [as can be seen from the scatterplot]. For a closer inspection of $\widehat{m}$ Figure 6 shows the influence of "amount" and "age" separately. In both plots of Figure 6 one variable is held fixed at levels 0.4 (short dashes), 0.5 (thick line) and 0.6 (long dashes). For "age" these levels correspond

	Coeff.	Std.Err.	$P > \|z\|$	Coeff.
const.	2.075	0.616	0.0001	—
previous credits o.k.	-0.698	0.320	0.030	-0.763
employed	0.543	0.311	0.082	0.569
duration	-1.821	0.876	0.039	-2.248
amount	-1.002	1.014	0.324	—
age	0.821	0.688	0.234	—
	Linear (logit)			Part. Linear

Table 7: Logit coefficients and coefficient in partially linear fit for credit scoring, $n = 284$.

to 32.9, 37.75, and 43.30 years, respectively. For credit amounts the corresponding original values are DM 1735.90, DM 2463.46, and DM 3495.95, respectively. So obviously, a higher amount of credit seems to get more risky in conjunction with higher age. Also, younger people seem to get less risky with increasing credit amount. Both of these possible conclusions could not be seen from the parametric logit fit.

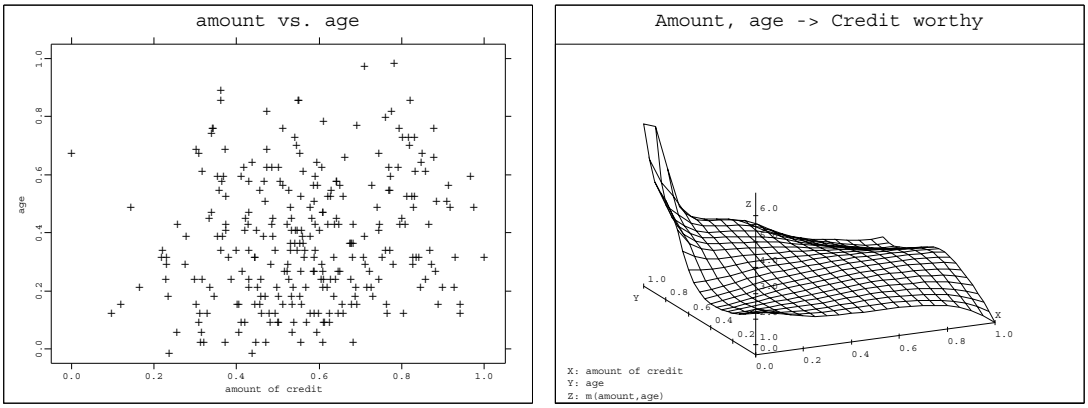

Figure 5: Scatterplot for amount of credit and age (left panel). Influence $\widehat{m}(t_1, t_2)$ of amount and age on credit worthiness (right panel), $n = 284$.

Table 7 gives the observed significance levels of our test statistics for the credit data. For the tests R_1 and R_2 linearity is rejected at level 0.10 for $h \leq 0.5$. For

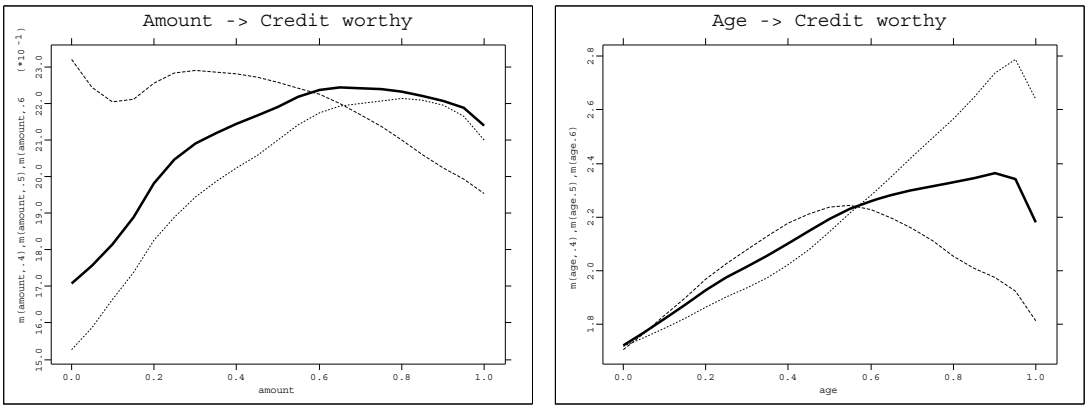

Figure 6: Influence of amount on credit worthiness for fixed age (left panel). Influence of age on credit worthiness for fixed amount (right panel). $n = 284$.

$h = 0.2$ the rejection has even higher significance. This suggests that deviances from linearity are more locally concentrated. Our inference in both applications was based on inspection of several tests. For getting a resulting p-value one could consider a combination of the test statistics for several bandwidths and one could calculate critical values for this combined statistic again by bootstrap.

h	0.2	0.3	0.4	0.5	0.6
R_1	0.04	0.08	0.08	0.09	0.28
R_2	0.02	0.07	0.07	0.09	0.29
R_3	0.24	0.12	0.07	0.08	0.27

Table 8: Observed significance levels for linearity test for credit scoring, $n = 284$. 400 bootstrap replications.

A1 Computational Remarks

In this section we indicate how the estimates in (2.3) and (2.4) can be numerically computed. The following algorithm corresponds to that proposed in Severini and Staniswalis (1994), Example 3, for the special case of a logistic link function.

Put $\eta_j(\beta) = \widehat{m}_\beta(T_j)$ and

$$L_i(u) = Q\{G(u); Y_i\}. \tag{A1.1}$$

Note, that we have

$$\begin{aligned} L_i'(u) &= \frac{Y_i - G(u)}{V(G(u))\}} G'(u) && \text{(A1.2)} \\ L_i''(u) &= \{Y_i - G(u)\} \left[\frac{G''(u)}{V(G(u))} \right. \\ &\qquad \left. - \frac{V'(G(u))\ G'(u)^2}{V(G(u))^2} \right] - \frac{G'(u)^2}{V(G(u))}. && \text{(A1.3)} \end{aligned}$$

Then maximizing the smoothed quasi–likelihood (2.2) requires to solve

$$0 = \sum_{i=1}^{n} L_i'\{X_i^T\beta + \eta_j(\beta)\} K_h(T_i - T_j). \tag{A1.4}$$

Differentiation of (A1.4) leads to $0 = \sum\limits_{i=1}^{n} L_i''\{X_i^T\beta + \eta_j(\beta)\} K_h(T_i - T_j)\{X_i + \eta_j'(\beta)\}$. This gives

$$\eta_j'(\beta) = \frac{-\sum\limits_{i=1}^{n} L_i''\{X_i^T\beta + \eta_j(\beta)\} K_h(T_i - T_j) X_i}{\sum\limits_{i=1}^{n} L_i''\{X_i^T\beta + \eta_j(\beta)\} K_h(T_i - T_j)}. \tag{A1.5}$$

For $\beta = \widehat{\beta}$ it holds that

$$0 = \sum_{i=1}^{n} L_i'\{X_i^T\beta + \eta_i(\beta)\}\ \{X_i + \eta_i'(\beta)\}. \tag{A1.6}$$

Equations (A1.4), (A1.5), (A1.6) suggest the following iterative Newton–Raphson type algorithm to find $\widehat{\beta}$ and $\widehat{m}(T_j)$, $j = 1, \ldots, n$.

- Start with $\widehat{\beta}^0 = \widetilde{\beta}$, $\widehat{\eta}_j^0 = T_j^T\widetilde{\gamma}$.
- The iteration $k \to k+1$ is determined by the stepwise application of the following two equations:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} L_i'(X_i^T\widehat{\beta}^k + \widehat{\eta}_j^k) K_h(T_i - T_j) + L_i''(X_i^T\widehat{\beta}^k + \widehat{\eta}_j^k) K_h(T_i - T_j)(\widehat{\eta}_j^{k+1} - \widehat{\eta}_j^k) \\ 0 &= \sum_{i=1}^{n} L_i'(X_i^T\widehat{\beta}^k + \widehat{\eta}_i^{k+1}) \widetilde{X}_i^k + L_i''(X_i^T\widehat{\beta}^k + \widehat{\eta}_i^{k+1}) \widetilde{X}_i^k \widetilde{X}_i^{kT} (\widetilde{\beta}^{k+1} - \widetilde{\beta}^k), \end{aligned}$$

where

$$\widetilde{X}_j^k = X_j - \frac{\sum_{i=1}^n L_i''(X_i^T\widehat{\beta}^k + \widehat{\eta}_j^{k+1})K_h(T_i - T_j)X_i}{\sum_{i=1}^n L_i''(X_i^T\widehat{\beta}^k + \widehat{\eta}_j^{k+1})K_h(T_i - T_j)}.$$

Then $\widehat{m}^k(T_j) = \widehat{\eta}_j^k$.

Alternatively, the functions $L_i''(u)$ can be replaced by their expectations $-G'(u)^2/V\{G(u)\}$ to obtain a Fisher scoring type procedure.

A2 Assumptions

We state now the assumptions used in the results in Section 3. In the following, the underlying parameters are denoted by β_0, γ_0 and m_0. We use the notation

$$h_{max} = max\{h_1, \ldots, h_q\},$$

$$h_{prod} = h_1 \cdot \ldots \cdot h_q,$$

$$\rho = h_{max}^2 + (nh_{prod})^{-1/2},$$

$$\tau = h_{max} + (nh_{prod})^{-1/2}.$$

For the asymptotic expansions we make the following assumptions.

(A1) $(X_1, T_1, Y_1), \ldots, (X_n, T_n, Y_n)$ are i.i.d. tuples with values in $\mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}$.

(A2) $E(Y_i|X_i, T_i) = G\{X_i^T\beta_0 + m_0(T_i)\}$ with $\beta_0 \in \mathbb{R}^p$. The conditional variance $Var(Y_i|T_i = t)$ has a bounded second derivative. Furthermore the Laplace transform $E\ \exp t|Y_i|$ is finite for $t > 0$ small enough.

(A3) $X_i^T\beta_0 + m_0(T_i)$ has compact support S. X_i and T_i have compact convex support S_X, S_T. T_i has a twice continuously differentiable density f_T with $\inf_{t \in S_T} f_T(t) > 0$.

(A4) There exists an $\delta > 0$ such that $G^{(k)}(u), \quad k = 1, \ldots, 3$ and $G'(u)^{-1}$ are bounded on $u \in S^\delta = \{v : \exists v' \in S \text{ with } |v' - v| \leq \delta\}$. Furthermore V^{-1}, V' and V'' are bounded on $G(S^\delta)$.

(A5) The kernel K is a product kernel $K(u) = K_1(u_1) \cdot \ldots \cdot K_q(u_q)$. The kernels K_j are symmetric probability densities with compact support ($[-1, 1]$, say), $j = 1, \ldots, q$.

(A6) The estimate $\widehat{\beta}$ is defined as $\arg \max_{\beta: \|\beta - \beta_0\| \leq \rho} \mathcal{L}(\widehat{m}_\beta, \beta)$. For a δ_n with $\delta_n \to 0$ the estimate $\widehat{m}_\beta(t)$ is defined as $\arg \max_{\eta: |\eta - m_0(t)| \leq \delta_n} \sum_{i=1}^{n} L_i(X_i^T \beta + \eta) K_h(T_i - t)$.

(A7) $E\left[L_1''\{X_1^T \beta_0 + m_0(T_1)\} | T_1 = t\right]$ and $E\left[L_1''\{X_1^T \beta_0 + m_0(T_1)\} X_1 | T_1 = t\right]$ are twice continuously differentiable functions for $t \in S_T$.

(A8) $h_{prod}\, n^{1/2} (\log n)^{-1} \to \infty$ and $h_{max} = o(n^{-1/8} (\log n)^{-1/4})$.

A3 Proofs

In this section we always assume that (A1) - (A7) hold. The following lemmas give the stochastic expansions for $\widehat{\beta}$ and $\widehat{m}$. Recall that the set S_T was the (compact) support of T_i. We denote $S_T^- = \{t \in S_T : t + \eta \in S_T \text{ for all } \eta \text{ with } |\eta_j| \leq h_j (j = 1, \ldots, q)\}$ and $S_T^h = S_T \setminus S_T^-$. Furthermore, define

$$\begin{aligned} S_{i,1} &= L_i'\{X_i^T \beta_0 + m_0(T_i)\}, \quad S_{i,2} = L_i''\{X_i^T \beta_0 + m_0(T_i)\}, \\ \widetilde{X}_i &= X_i - \{E[S_{i,2} | T_i]\}^{-1} E[S_{i,2} X_i | T_i], \\ w_i(t) &= K_h(t - T_i) \Big\{ n^{-1} \sum_{j=1}^{n} K_h(t - T_j) \Big\}^{-1}. \end{aligned}$$

Lemma A3.1

(i) For all $C > 0$ it holds that

$$\sup_{\substack{t \in S_T^- \\ \|\beta - \beta_0\| \leq C\rho}} \Bigg| \widehat{m}_\beta(t) - \Bigg(m(t) - \{E(S_{1,2} | T_1 = t)\}^{-1} \Big[\frac{1}{n} \sum_{i=1}^{n} w_i(t) L_i'\{X_i^T \beta_0 + m_0(t)\} + E(S_{1,2} X_1^T | T_1 = t)(\beta - \beta_0) \Big] \Bigg) \Bigg| = O_p(\rho^2 \log n).$$

(ii) The supremum in (i) taken over $t \in S_T^h, \|\beta - \beta_0\| \leq C\rho$ *is of stochastic order* $O_p(\tau^2)$.

Proof

We prove only statement (i). Choose $C > 0$. We have for $t \in S_T^-, \|\beta - \beta_0\| \leq C\rho$

$$\sum_{i=1}^{n} L_i'\{X_i^T\beta + \widehat{m}_\beta(t)\}K_h(t - T_i) = 0. \tag{A3.1}$$

This follows from

$$\sup \sum_{i=1}^{n} L_i''(X_i^T\beta + \eta)K_h(t - T_i) < 0 \tag{A3.2}$$

with probability tending to one, where the supremum runs over $|\eta - m_0(t)| \leq \delta_n, t \in S_T^-$, and β with $\|\beta - \beta_0\| \leq C\rho$.

Note that (A3.2) implies that, if we find an $\eta_\beta(t)$ with $|\eta_\beta(t) - m_0(t)| \leq \delta_n$ and

$$\sum_{i=1}^{n} L_i'\{X_i^T\beta + \eta_\beta(t)\}K_h(t - T_i) = 0,$$

then with probability tending (uniformly) to one we get $\widehat{m}_\beta(t) = \eta_\beta(t)$. Inequality (A3.2) can be shown by using that for $\delta > 0$ small enough

$$sup_{\eta \in I_n', \beta \in I_n'', t \in I'''} \left| \frac{1}{n} \sum_{i=1}^{n} L_i''\{X_i^T\beta + \eta\}K_h(t - T_i) \right.$$
$$\left. - E\left[L_i''\{X_i^T\beta + \eta\}K_h(t - T_i)\right] \right| = o_P(1) \tag{A3.3}$$

$$\sup_{1 \leq i \leq n} \sup_{u \in S^\delta, t \in I\!R^q} |L_i'''(u)K_h(t - T_i)| = O_P(h_{prod}^{-1} \log n), \tag{A3.4}$$

$$\sup_{1 \leq i \leq n} \sup_{u \in S^\delta, t \in I\!R^q} \|L_i'''(u)K_h'(t - T_i)\| = O_P(h_{prod}^{-1} h_{max}^{-1} \log n), \tag{A3.5}$$

where the supremum in (A3.3) runs over grids I', I'' and I''' with polynomially many elements. Equality (A3.3) follows by application of the Markov inequality. Note that Y_i has bounded Laplace transform, see Assumption (A2). Equalities (A3.4) - (A3.5) follow from $\max_{1 \leq i \leq n} |Y_i| = O_P(\log n)$. This can be shown again by using that Y_i has bounded Laplace transform. For the proof of claim (A3.2) one applies

$$E\left[L_i''\{X_i^T\beta + \eta\}K_h(t - T_i)\right] = -E\frac{G'\{X_i^T\beta + \eta\}^2}{V[G\{X_i^T\beta + \eta\}]}.$$

Equation (A3.1) implies

$$
\begin{aligned}
0 \;=\; & \frac{1}{n}\sum_{i=1}^{n} w_i(t) L_i'\{X_i^T\beta_0 + m_0(t)\} \qquad \text{(A3.6)}\\
& + \frac{1}{n}\sum_{i=1}^{n} w_i(t) L_i''\{X_i^T\beta_0 + m_0(t)\}\,\{\widehat{m}_\beta(t) - m_0(t) + X_i^T(\beta-\beta_0)\}\\
& + R_1(\beta,t)\left[\{\widehat{m}_\beta(t) - m_0(t)\}^2 + \rho^2\right]
\end{aligned}
$$

with

$$
\sup_{\substack{t\in S_T^- \\ \|\beta-\beta_0\|\le C\rho}} |R_1(\beta,t)| \le C_1 \quad \text{(a.s)}
$$

for a constant $C_1 > 0$ for n large enough. Furthermore, we have $|\widehat{m}_\beta(t) - m_0(t)| \le \delta_n \to 0$, see (A6). This implies

$$
\begin{aligned}
\widehat{m}_\beta(t) \;=\; & m_0(t) - \left[\frac{1}{n}\sum_{i=1}^{n} w_i(t) L_i''\{X_i^T\beta_0 + m_0(t)\}\right]^{-1} \qquad \text{(A3.7)}\\
& \left[\frac{1}{n}\sum_{i=1}^{n} w_i(t) L_i'\{X_i^T\beta_0 + m_0(t)\} + \frac{1}{n}\sum_{i=1}^{n} w_i(t) L_i''\{X_i^T\beta_0 + m_0(t)\} X_i^T(\beta-\beta_0)\right]\\
& + R_2(\beta,t)\,\rho^2 \log n,
\end{aligned}
$$

where

$$
\sup_{\substack{t\in S_T^- \\ \|\beta-\beta_0\|\le C\rho}} |R_2(\beta,t)| = O_p(1).
$$

For (A3.7) it has been used that

$$
\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^{n} w_i(t) L_i'\{X_i^T\beta_0 + m_0(t)\}\right| = O_p(\rho\sqrt{\log n}).
$$

This follows from

$$
\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^{n} K_h(t-T_i)\left[L_i'\{X_i^T\beta_0 + m_0(t)\} - L_i'\{X_i^T\beta_0 + m_0(T_i)\}\right]\right| = O_p(\rho)
$$

and

$$
\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^{n} K_h(t-T_i) L_i'\{X_i^T\beta_0 + m_0(T_i)\}\right| = O_p(\rho\sqrt{\log n}).
$$

Recall that $E\left[L_i'\{X_i^T\beta_0 + m_0(T_i)\}|X_i, T_i\right] = 0$. For the statement of the lemma it remains to show

$$\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^n w_i(t)L_i''\{X_i^T\beta_0 + m_0(t)\} - E(S_{1,2}|T_1 = t)\right| \tag{A3.8}$$
$$= O_p(\rho\sqrt{\log n})$$
$$\sup_{t\in S_T^-}\left\|\frac{1}{n}\sum_{i=1}^n w_i(t)L_i''\{X_i^T\beta_0 + m_0(t)\}X_i^T - E(S_{1,2}X_1^T|T_1 = t)\right\| \tag{A3.9}$$
$$= O_p(\rho\sqrt{\log n}).$$

For the proof of (A3.8) note first that

$$\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^n w_i(t)\left[L_i''\{X_i^T\beta_0 + m_0(t)\} - L_i''\{X_i^T\beta_0 + m_0(T_i)\}\right]\right| = O_p(\rho),$$

see (A4). With the help of (A7) one shows

$$\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^n w_i(t)L_i''\{X_i^T\beta_0 + m_0(T_i)\} - E(S_{1,2}|T_1 = t)\right| = O_p(\rho\sqrt{\log n}).$$

Equation (A3.9) can be shown similarly.

Lemma A3.2

(i) For all $C > 0$ it holds that

$$\sup_{\substack{t\in S_T^- \\ \|\beta-\beta_0\|\le C\rho}}\left\|\frac{\partial\widehat{m}_\beta(t)}{\partial\beta} + \{E(S_{1,2}|T_1 = t)\}^{-1}E(S_{1,2}X_1|T_1 = t)\right\| = O_p(\rho\sqrt{\log n}).$$

(ii) The supremum in (i) taken over $t \in S_T^h, \|\beta - \beta_0\| \le C\rho$ is of stochastic order $O_p(\tau)$.

Proof

Lemma A3.2 can be proved similarly as Lemma A3.1. One uses that

$$\sum_{i=1}^n L_i''\{X_i^T\beta + \widehat{m}_\beta(t)\}K_h(t - T_i)\frac{\partial}{\partial\beta}\widehat{m}_\beta(t) \tag{A3.10}$$
$$+\sum_{i=1}^n L_i''\{X_i^T\beta + \widehat{m}_\beta(t)\}X_iK_h(t - T_i) = 0.$$

Lemma A3.3
For the estimate $\widehat{\beta}$ the following stochastic expansion holds

$$\widehat{\beta} = \beta_0 + \{E(S_{1,2}\widetilde{X}_1\widetilde{X}_1^T)\}^{-1}\frac{1}{n}\sum_{i=1}^{n} S_{i,1}\widetilde{X}_i + O_p(\rho^2 \log n).$$

Proof
We show that with probability tending to one there exists a solution β with $\|\beta - \beta_0\| \leq \rho$ of the following equation and that (with probability tending to one) this solution is unique.

$$\frac{\partial}{\partial\beta}\sum_{i=1}^{n} L_i\{X_i^T\beta + \widehat{m}_\beta(T_i)\} = 0. \tag{A3.11}$$

Expansion of the left hand side of (A3.11) gives with the help of Lemma A3.2

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} L_i'\{X_i^T\beta + \widehat{m}_\beta(T_i)\}\left[X_i + \frac{\partial}{\partial\beta}\widehat{m}_\beta(T_i)\right] \qquad (A3.12)\\ &= \frac{1}{n}\sum_{i=1}^{n} L_i'\{X_i^T\beta_0 + m_0(T_i)\}\left[X_i + \frac{\partial}{\partial\beta}\widehat{m}_\beta(T_i)\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} L_i''\{X_i^T\beta_0 + m_0(T_i)\}\widetilde{X}_i X_i^T(\beta - \beta_0) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} L_i''\{X_i^T\beta_0 + m_0(T_i)\}\widetilde{X}_i[\widehat{m}_\beta(T_i) - m_0(T_i)] + O_p(\rho^2 \log n). \end{aligned}$$

This expansion holds uniformly for β with $\|\beta - \beta_o\| \leq \rho$. For instance, it has been used that

$$\sup_{\substack{t\in S_t^- \\ \|\beta-\beta_0\|\leq\rho}} |\widehat{m}_\beta(t) - m(t)| = O_p(\rho\sqrt{\log n}).$$

This follows by standard techniques from Lemma A3.1. By expansion of (A3.10) it can be shown that

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} L_i'\{X_i\beta_0 + m_0(T_i)\}\left[X_i + \frac{\partial}{\partial\beta}\widehat{m}_\beta(T_i)\right] \\ &\qquad = \frac{1}{n}\sum_{i=1}^{n} L_i'\{X_i^T\beta_0 + m_0(T_i)\}\widetilde{X}_i + O_p(\rho^2). \end{aligned}$$

Plugging this into the right hand side of (A3.12) and replacing averages by their expectations gives that (with probability tending to one) there exists a solution

$\beta = \overline{\beta}$ of (A3.11) with

$$\overline{\beta} = \beta_0 + \{E(S_{1,2}\widetilde{X}_1\widetilde{X}_1^T)\}^{-1}\,\frac{1}{n}\sum_{i=1}^{n} S_{i,1}\widetilde{X}_i + \; O_p(\rho^2 \log n).$$

Because of $\overline{\beta} - \beta_0 = O_p(n^{-1/2})$, we have $\overline{\beta} = \widehat{\beta}$ (with probability tending to one). This shows Lemma A3.3.

With the help of Lemmas A3.1 and A3.2 we get for the estimate $\widehat{m}$ the following expansion.

Corollary A3.4

(i) For the estimate $\widehat{m}$ the following stochastic expansion holds:

$$\sup_{t\in S_T^-}\Big|\widehat{m}(t) - \{\overline{m}(t) + \{E(S_{1,2}|T_1=t)\}^{-1}E(S_{1,2}X_1^T|T_1=t)$$
$$\{(S_{1,2}\widetilde{X}_1\widetilde{X}_1)\}^{-1}\frac{1}{n}\sum_{i=1}^{n} S_{i,1}\widetilde{X}_i\}\Big| = \; O_p(\rho^2\sqrt{\log n}),$$

with $\overline{m}(t) = m_0(t) + E(S_{1,2}|T_1=t)^{-1}\frac{1}{n}\sum_{i=1}^{n} w_i(t)L_i'\{X_i^T\beta_0 + m_0(t)\}$.

(ii) The supremum in (i) taken over $t \in S_T^h$ is of stochastic order $O_p(\tau^2)$.

In particular, we get $\sup_{t\in S_T^-}|\widehat{m}(t) - \overline{m}(t)| = O_p(n^{-1/2})$ and $\sup_{t\in S_T^h}|\widehat{m}(t) - \overline{m}(t)| = O_p(\tau^2)$. Also $\sup_{t\in S_T^-}|\widehat{m}(t) - m(t)| = O_p(\rho\sqrt{\log n})$ and $\sup_{t\in S_T^h}|\widehat{m}(t) - m(t)| = O_p(\tau)$.
In Section 2 we introduced in (2.9) the modification $\widetilde{m}(t)$ of the parametric estimate $t^T\widetilde{\gamma}$. The purpose of this modification was to compensate for the bias of $\widehat{m}(t)$ when comparing $\widetilde{m}(t)$ and $\widehat{m}(t)$. The next lemma shows that this modification works.

Lemma A3.5
Suppose that the hypothesis (1.1) holds, i.e. $m_0(t) = t^T\gamma_0$.

$$\sup_{t\in S_T^-}\Big|\,\widetilde{m}(t) - t^T(\widetilde{\gamma} - \gamma_0) - E\{\overline{m}(t)|X_1, T_1, \ldots, X_n, T_n\}\Big| = \; O_p(\rho^2\sqrt{\log n}).$$

Proof

The proof uses similar expansions as above. In particular it uses the fact that with probability tending to one

$$\sum_{i=1}^{n} K_h(t-T_i)\frac{G\{\widetilde{\mu}_i(t)\}-G(X_i^T\widetilde{\beta}+T_i^T\widetilde{\gamma})}{G\{\widetilde{\mu}_i(t)\}[1-G\{\widetilde{\mu}_i(t)\}]}\,G'\{\widetilde{\mu}_i(t)\}=0,$$

where $\widetilde{\mu}_i(t)=X_i^T\widetilde{\beta}+\widetilde{m}(t)$.

Proof of Theorem 3.1

Application of the foregoing expansions for the parametric and semiparametric estimates gives:

$$\sup_{t\in S_T^-}\left|[\widehat{m}(t)-\widetilde{m}(t)]-[\overline{m}(t)-E\{\overline{m}(t)|X_1,T_1,\ldots,X_n,T_n\}]\right| = O_p(\rho^2\sqrt{\log n}),$$

$$\sup_{t\in S_T^-}\left|\overline{m}(t)-E\{\overline{m}(t)|X_1,T_1,\ldots,X_n,T_n\}\right| = O_p((nh_{prod})^{-1/2}\sqrt{\log n}),$$

These equalities together with the expansions for the suprema over S_T^- imply for $j=1,2,3$

$$\begin{aligned} R_j &= R+O_p(n\rho^2(nh_{prod})^{-1/2}\log n),\\ R &= \sum_{i=1}^{n}\frac{G'(\eta_i)^2}{G(\eta_i)\{1-G(\eta_i)\}}\left\{\overline{m}(T_i)-E\left[\overline{m}(T_i)|X_1,T_1,\ldots,X_n,T_n\right]\right\}^2, \end{aligned}$$

where $\eta_i=X_i^T\beta_0+T_i^T\gamma_0$ for $i=1,\ldots,n$. Under our assumptions, we have $n\rho^2$ $(nh_{prod})^{-1/2}\log n=o(h_{prod}^{-1/2})=o(v_n)$. This shows statement (i). For statement (ii) note that, conditionally given $X_1,T_1,\ldots,X_n,T_n$, the statistic R is a U-statistic. Proceeding as in Härdle and Mammen (1993) one can verify de Jong's (1987) conditions for asymptotic normality of U-statistics.

Proof of Theorem 3.2

As in the proof of Theorem 3.1 one shows for $j=1,2,3$ that

$$d_K\{R_j^*,\,N(e_n,v_n^2)\}\longrightarrow 0 \quad \text{(in probability)}. \tag{A3.13}$$

(Recall that e_n and v_n have been introduced in Theorem 3.1.) For this purpose one notes first that for all three versions of bootstrap $|Y_i^*|$ has bounded conditional Laplace transform [in a neighborhood of 0]. This has been shown in the proof of Theorem 5.1 in Mammen and van de Geer (1997). For the proof of (A3.13) one proceeds now as in the proof of Theorem 3.1.

References

Beran, R. (1986). Comment on "Jackknife, bootstrap, and other resampling methods in regression analysis" by C. F. J. Wu. *Annals of Statistics* **14**: 1295–1298.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion), *Annals of Statistics* **17**: 453–555.

Burda, M. (1993). The determinants of east–west german migration, *European Economic Review* **37**: 452–461.

Carroll, R., Fan, J., Gijbels, I. and Wand, M. (1995). Generalized partially single-index models, *Technical report*, Department of Statistics, Texas A&M University.

Chen, R., Härdle, W., Linton, O. and Severance-Lossin, E. (1996). Estimation and variable selection in additive nonparametric regression models, *in* W. Härdle and M. Schimek (eds), *Proceedings of the COMPSTAT Satellite Meeting Semmering 1994*, Physica Verlag, Heidelberg.

de Jong, P. (1987). A central limit theorem for generalized quadratic forms, *Probability Theory and Related Fields* **75**: 261–277.

Fahrmeir, L. and Hamerle, A. (1984). *Multivariate Statistische Verfahren*, De Gruyter, Berlin.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer.

Fan, J., Härdle, W. and Mammen, E. (1995). Direct estimation of low dimensional components in additive models, *Discussion paper*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models, *International Statistical Review* **55**: 245–259.

GSOEP (1991). *Das Sozio-ökonomische Panel (SOEP) im Jahre 1990/91*, Projektgruppe "Das Sozio-ökonomische Panel", Deutsches Institut für Wirtschaftsforschung. Vierteljahreshefte zur Wirtschaftsforschung, pp. 146–155.

Härdle, W., Huet, S., Mammen, E. and Sperlich, S. (1997). Semiparametric additive indices for binary response, *Technical report*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Härdle, W. and Mammen, E. (1993). Testing parametric versus nonparametric regression, *Annals of Statistics* **21**: 1926–1947.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I - III, *Math. Methods of Statist.* **2**: 85 – 114, 171 – 189, 249 – 268.

Lepski, O. V. and Spokoiny, V. G. (1995). Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative, unpublished manuscript.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika* **82**: 93–101.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*, Econometric Society Monographs No. 4, Cambridge University Press.

Mammen, E. (1992). *When does bootstrap work: asymptotic results and simulations*, Lecture Notes in Statist. **77**, Springer, Berlin.

Mammen, E. and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models, *Annals of Statistics* **25**: 1014–1035.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Vol. 37 of *Monographs on Statistics and Applied Probability*, 2 edn, Chapman and Hall, London.

Müller, M. (1997). Computer–assisted Generalized Partial Linear Models. to appear in *Interface'97 Proceedings*, Houston, Texas.

Robinson, P. M. (1988). Root n-consistent semiparametric regression, *Econometrica* **56**: 931–954.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models, *Journal of the American Statistical Association* **89**: 501–511.

Severini, T. A. and Wong, W. H. (1992). Generalized profile likelihood and conditionally parametric models, *Annals of Statistics* **20**: 1768–1802.

Speckman, P. E. (1983). Regression analysis for partially linear models, *Journal of the Royal Statistical Society, Series B* **50**: 413–436.

Tjøstheim, D. and Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: projections., *Journal of the American Statistical Association* **89**: 1398–1409.

Wu, C. F. J. (1986). Jackknife, bootstrap, and other resampling methods in regression analysis, *Annals of Statistics* **14**: 1261–1350.